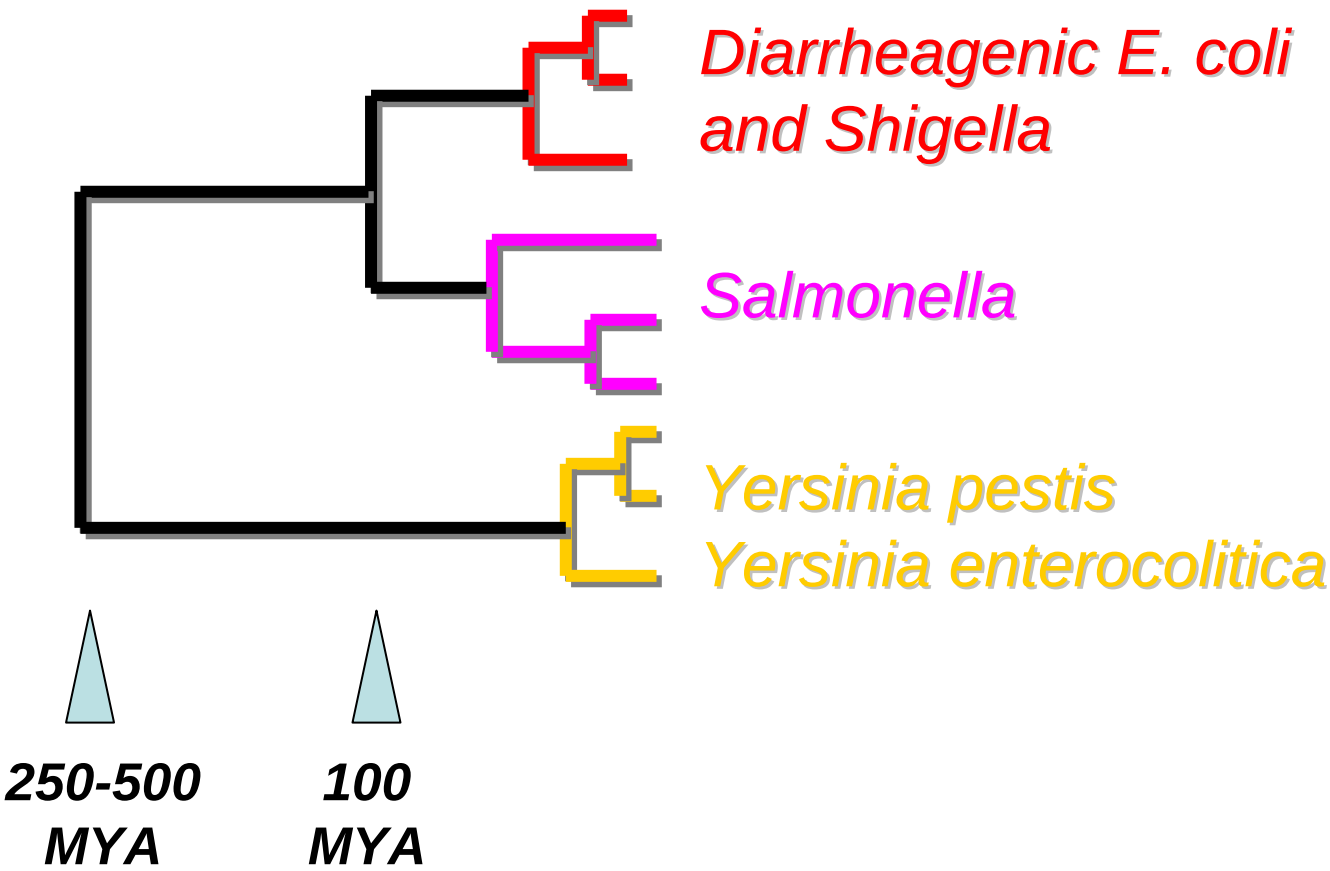
Mauve

Aaron Darling, Bob Mau, Paul Infield-Harm, Fred Blattner, Nicole Perna

ERIC Genomes are Closely Related



Sequence Alignment

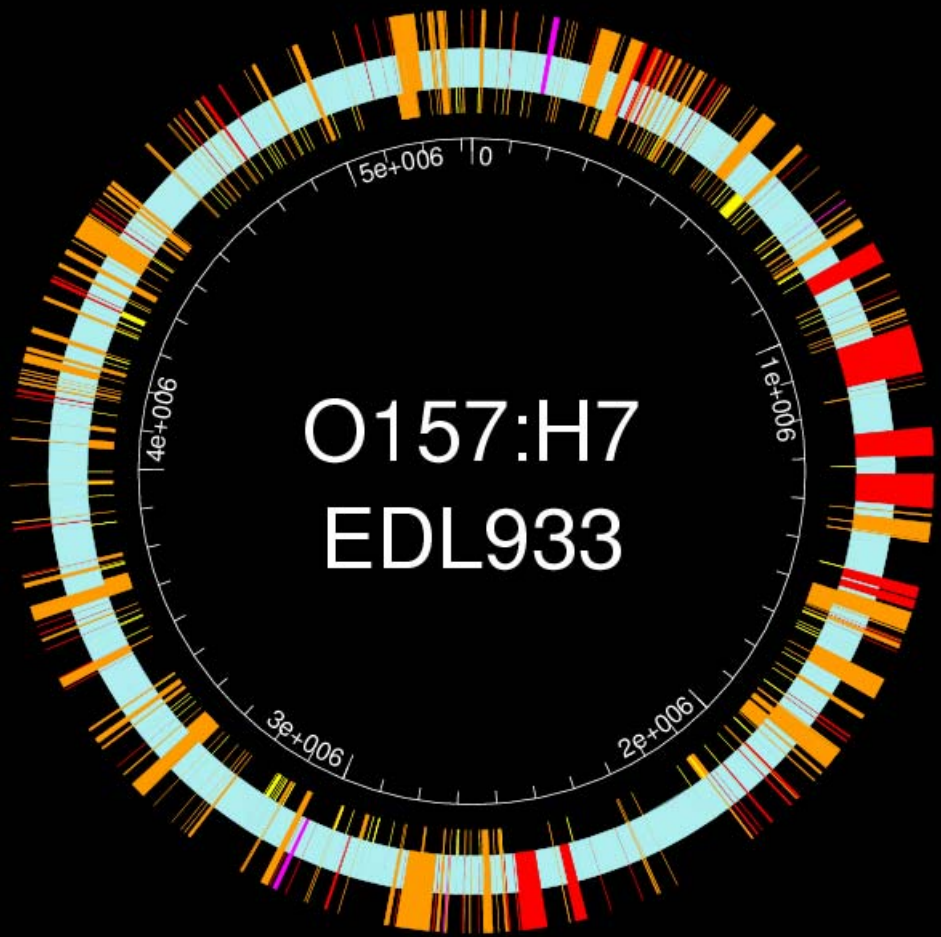
- Alignment is useful both to identify what is conserved…
  - Reusing annotations
  - Collinearity of putative orthologs
  - Identifying common targets for vaccines, etc.
  - Designing experiments
  - Evolution of pathogens
- And what isn't conserved
  - Organism/Group specific targets
  - Epidemiology

Much of the difference in gene content among enterobacterial genomes appears to be due to horizontal gene transfer offset by deletions.

E. coli O157:H7 EDL933 vs. K-12 MG1655

EDL933 islands = 1.5 Mb

MG1655 islands = 0.6 Mb

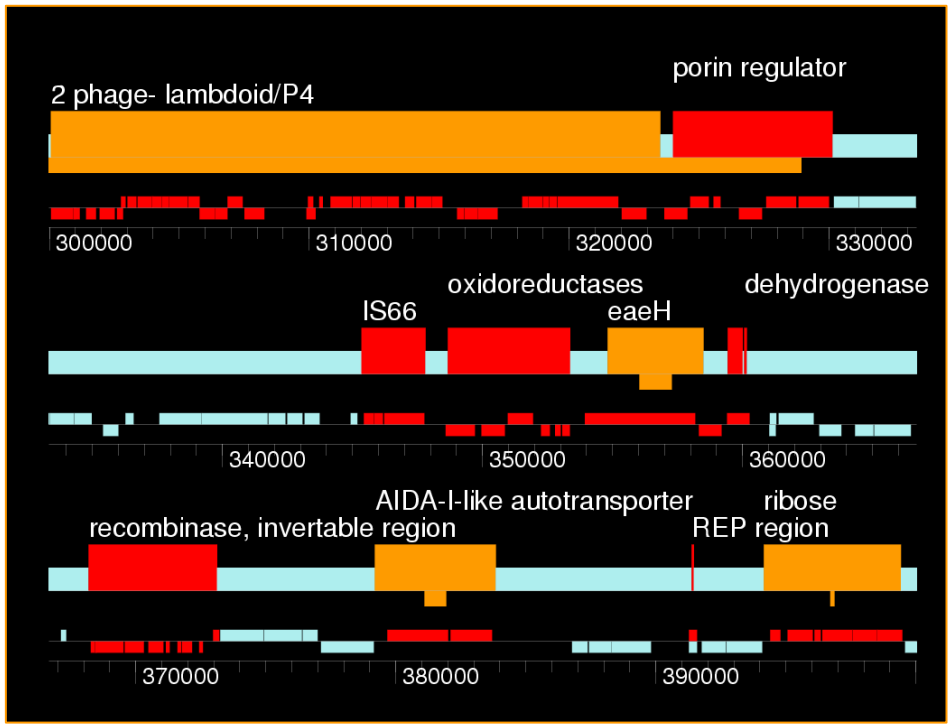


Backbone = 4.1 Mb

Hypervariable

Many genes associated with virulence are horizontally transferred elements.

polar fimbriae, other fimbriae, autotransporter family adhesins, invasins, other adhesins, RTX toxins, other toxins, iron compound transport, ribose transport, glutamate fermentation, urease, fatty acid biosynthesis, O-antigen biosynthesis, tellurite resistance, efflux pumps, sucrose utilization, proteases, type III secretion systems, insertion sequences, rhs elements, other outer membrane proteins,



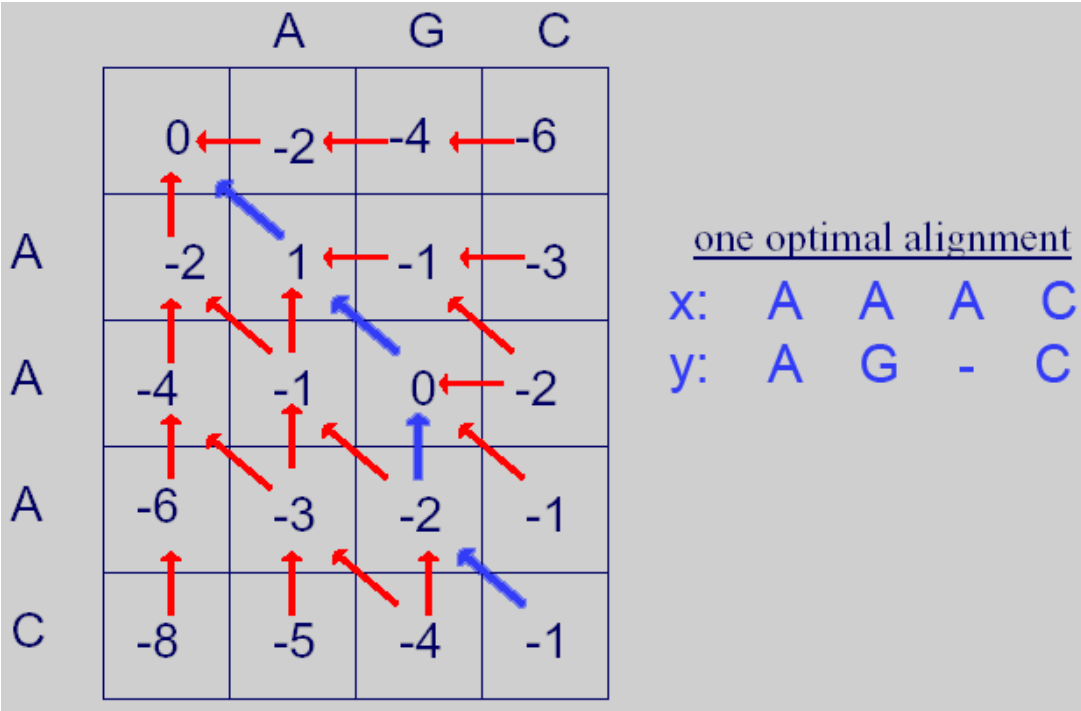
fatty acid degradation, xanthosine utilization, glyoxylate degradation, general protein secretion, galactonate degradation, cryptic beta-glucosidase, PTS, prophages, restriction/modification systems, non-ribosomal peptide synthesis, plant hormone biosynthesis, opine utilization, cell wall degegrading enzymes

Genome Evolution in Enterobacteria

- Differential gene content, even within species
- Co-localized clusters of differential content
- Genome rearrangement even within species
  - Two *E. coli*: 1 inversion
  - Two *Shigella*: at least 15 collinear segments
  - Two *Yersinia*: at least 25 collinear segments

Traditional pairwise global alignment

Scales $O(n^2)$ where n is the sequence length

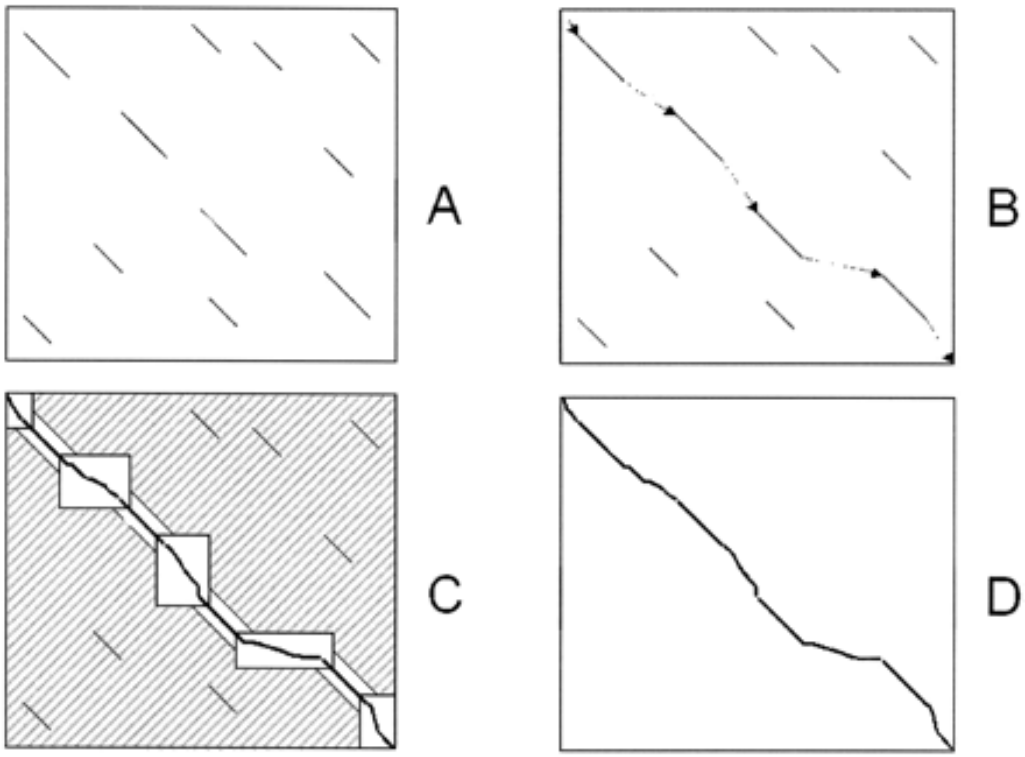


Multiple alignment scales $O(n^m)$ where m is the number of Sequences.

Problem: time-consuming, doesn't consider rearrangements

Anchored Alignment

Restrict the search to parts of the DP matrix that are very likely to be part of the optimal path



Each diagonal ‘band’ is a high-scoring local alignment of the sequences.

The highest scoring chain of local alignments become anchors

Anchored genome alignment tools

Multi-LAGAN – align two or more (divergent) genomes, assuming no differential gene content and no rearrangements (Brudno et. al. 2003)

MAVID – Like Multi-LAGAN, but also infer the branching structure of the organism's phylogeny (Bray et. al. 2004)

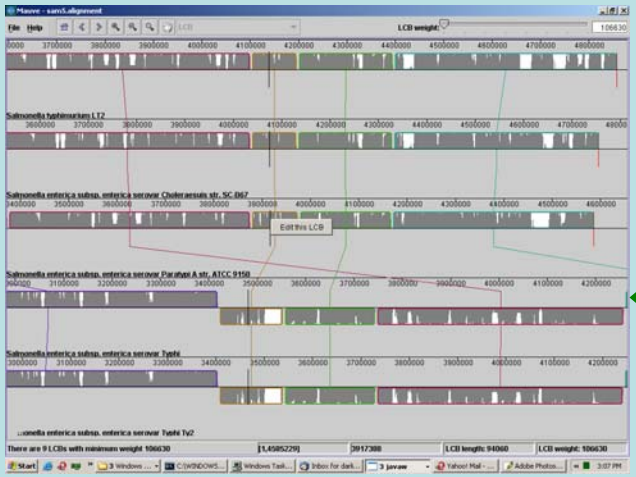
Shuffle-LAGAN – align two genomes that may contain repeats and rearrangements, no differential gene content (Brudno et. al. 2003)

Mauve – align two or more closely related genomes that have rearrangements, differential content in conserved order and orientation (Darling et. al. 2004)

Mulan – align two or more closely related genomes, possibly with differential gene content (Ovcharenko et. al. 2005)

The two component architecture of Mauve

Java 1.4 Interactive Visualization

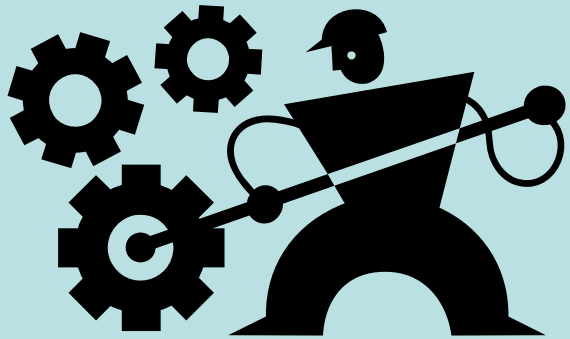


GenBank or FastA sequences

alignments

C++ command-line aligner

Windows, Linux, Mac OS X



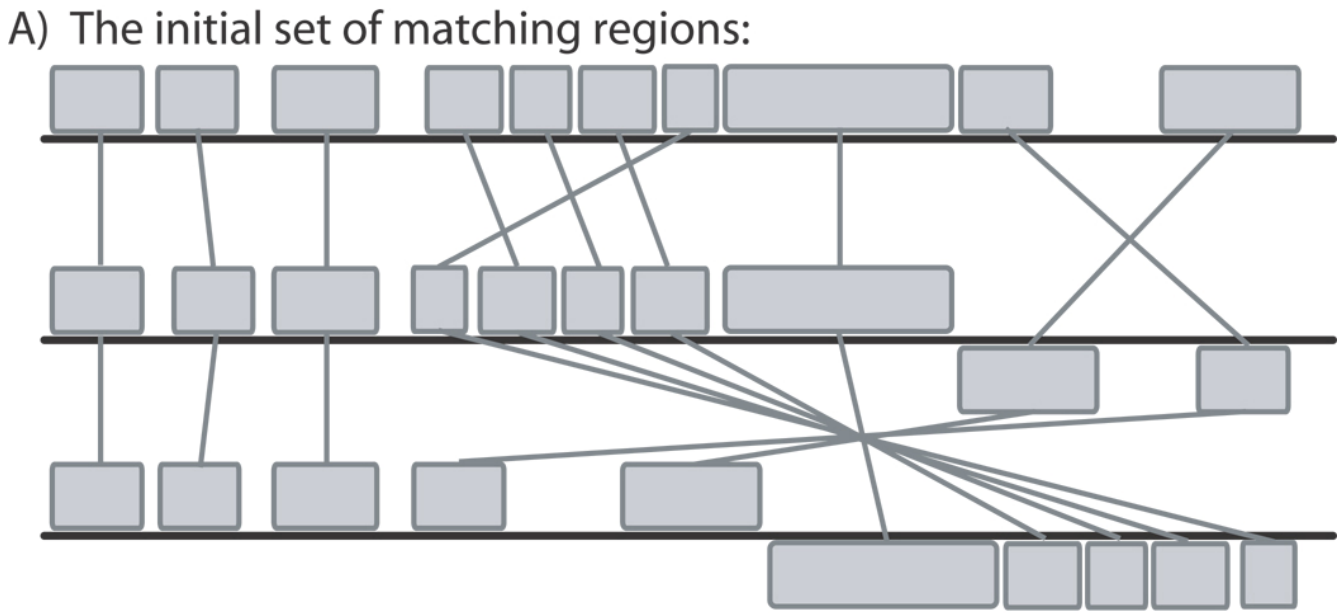
100% Free/Open Source Software

We use each language for what it does best–C++ for efficient algorithm implementation, Java for a cross platform GUI

The Mauve alignment approach

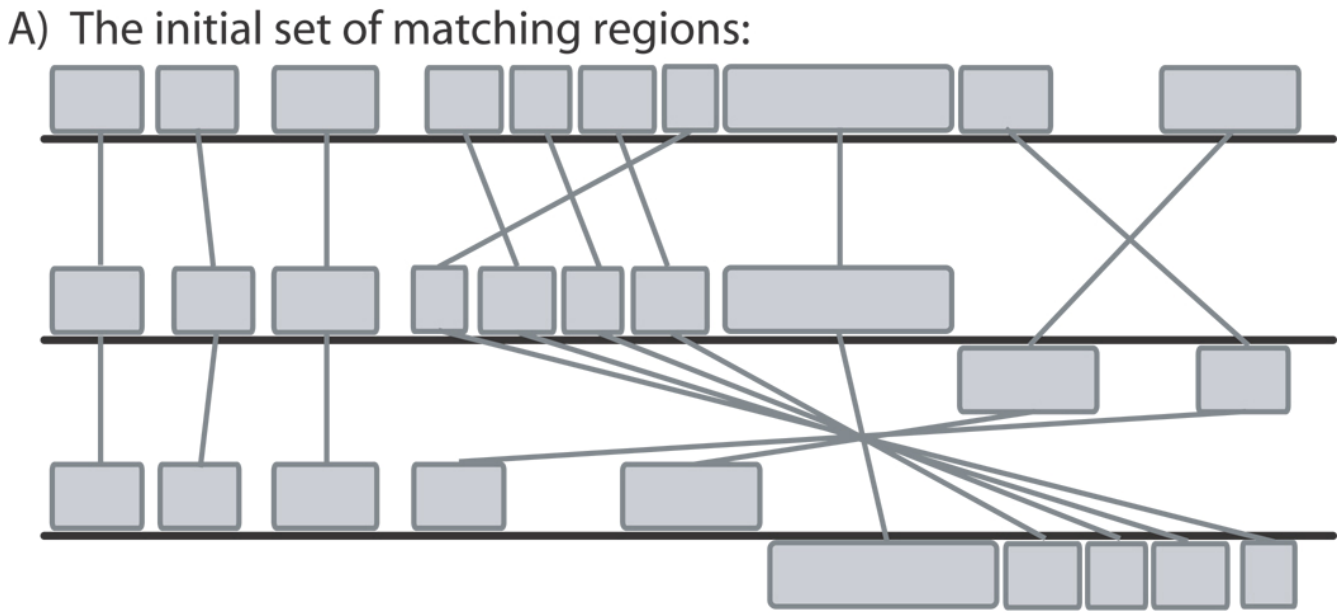
Each set of linked boxes is a high-scoring local alignment.

Boxes below a genome's center line are in the reverse-complement orientation (inverted)



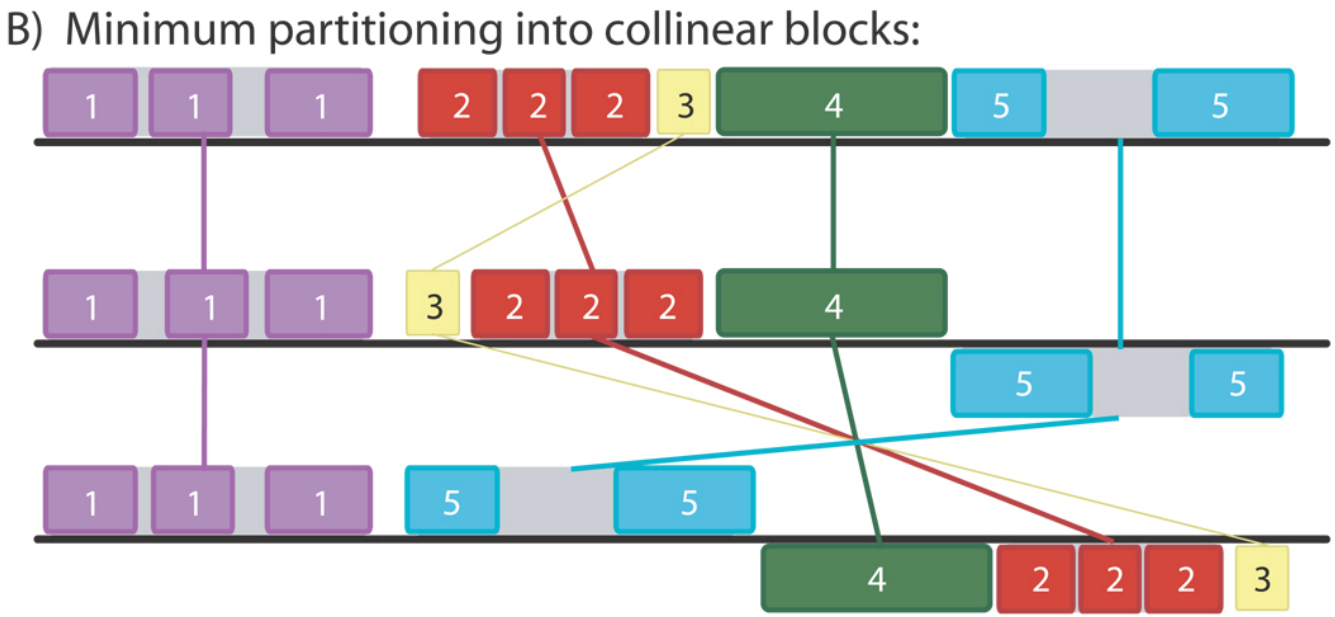
The Mauve alignment approach

Need to filter out matches that arise due to random sequence similarity (or paralogy)



The Mauve alignment approach

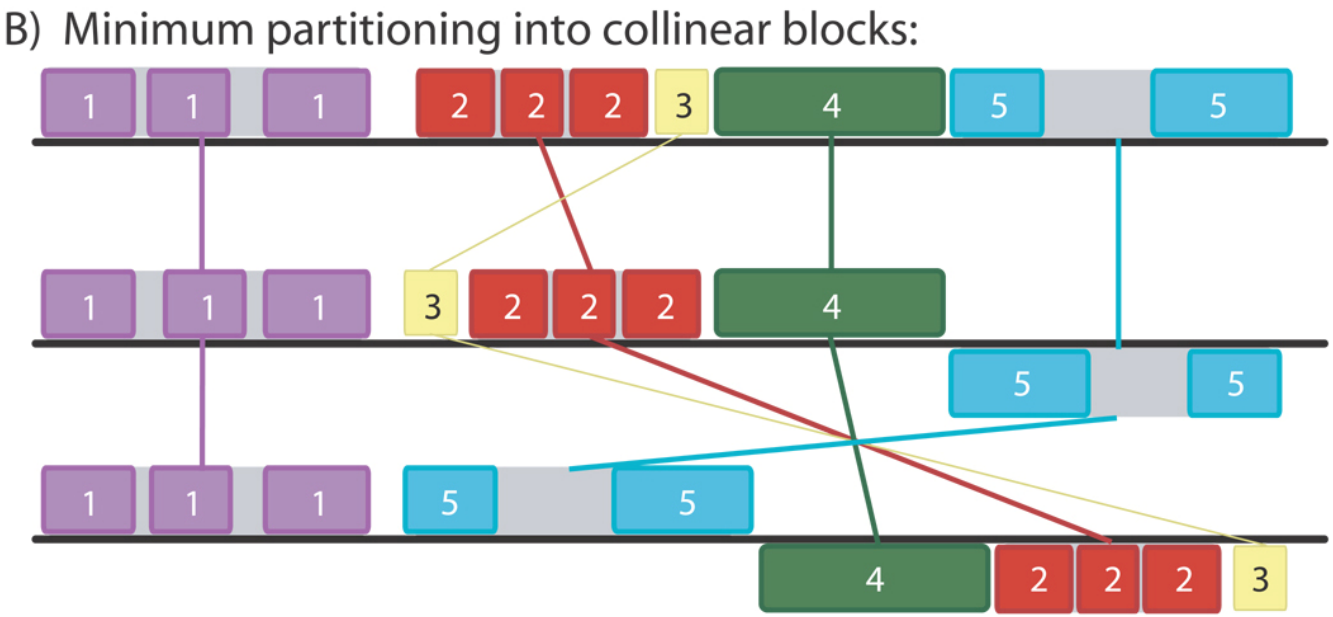
Use breakpoint analysis to identify Locally Collinear Blocks – groups of anchors with conserved order and orientation



Greedy Breakpoint Elimination

Remove matches caused by random similarity:

Block 3 (yellow) is small and has a weight less than w so it is removed.

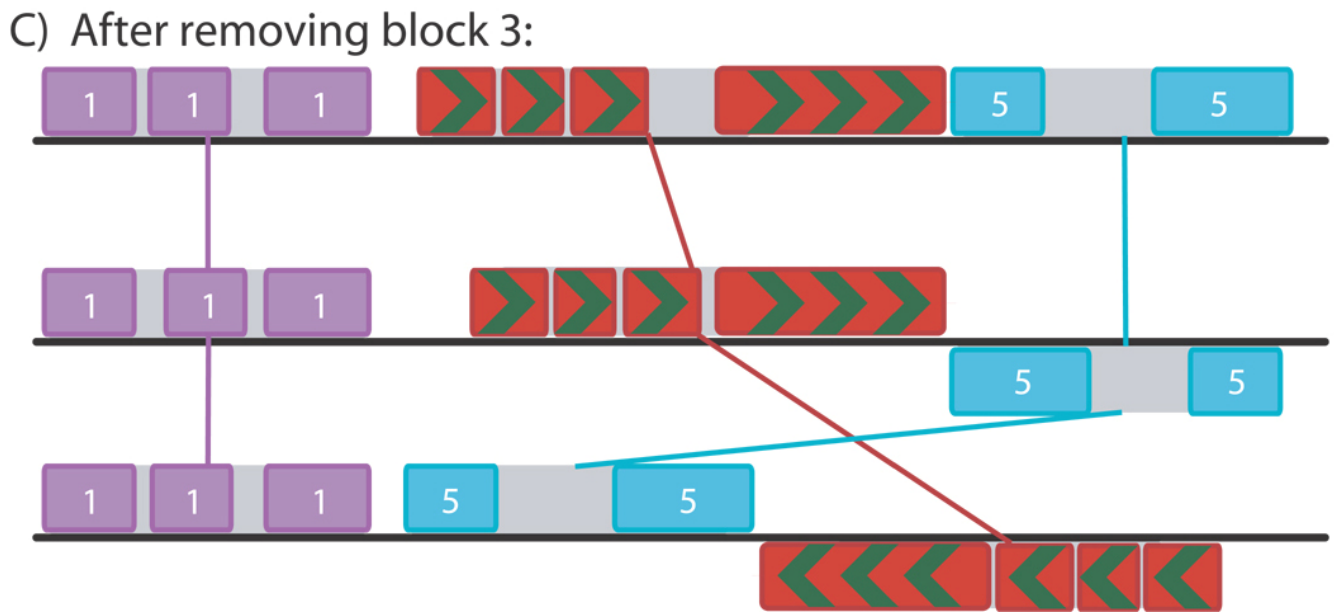


Greedy Breakpoint Elimination

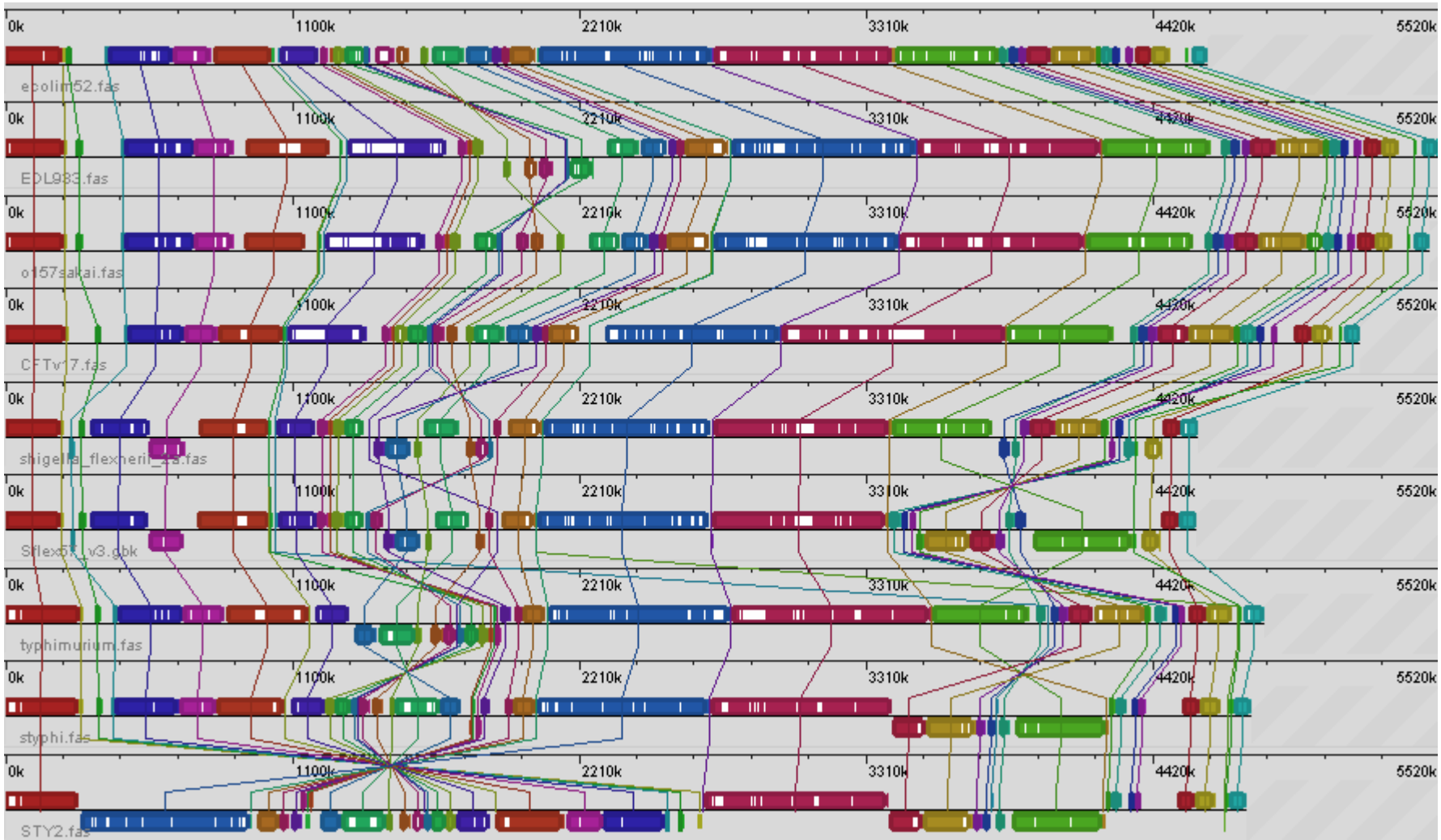
A breakpoint is eliminated:

When block 3 is removed, blocks 2 and 4 coalesce.

Final step: align grey regions progressively (with MUSCLE or Clustal-W)

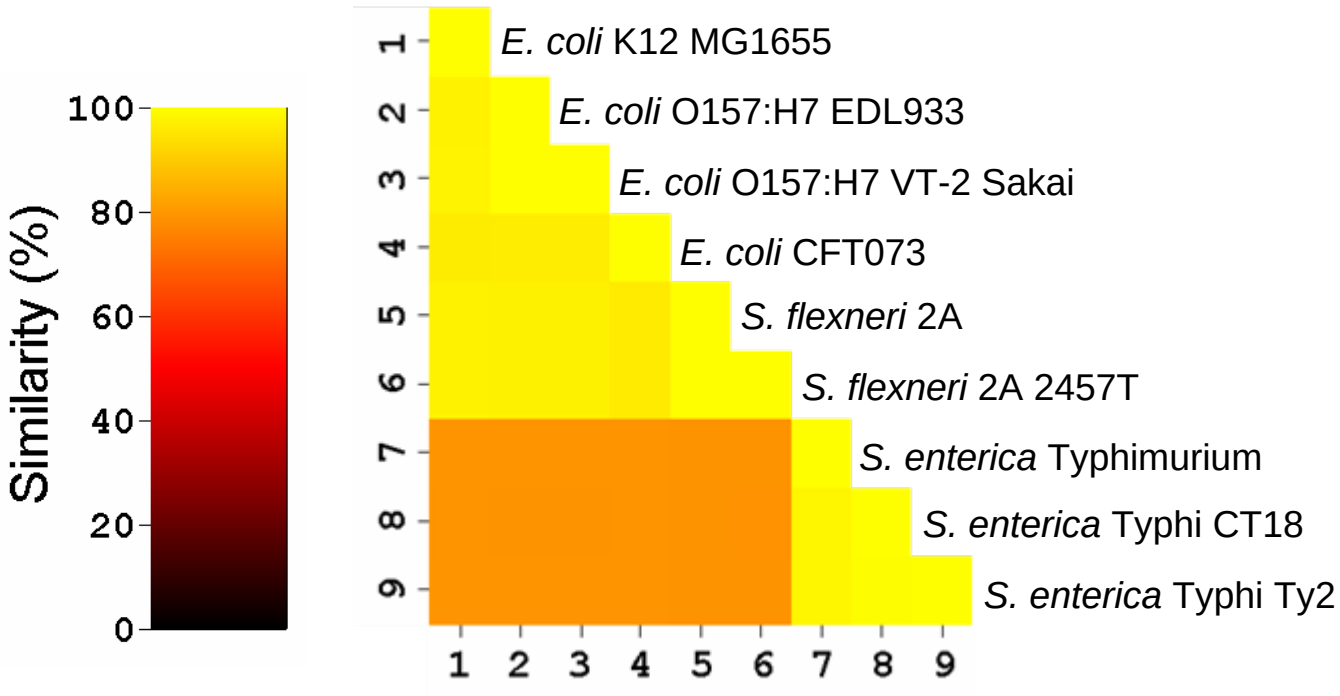


Nine Enterobacteria



Alignment of 9 Enterobacteria

- 45 locally collinear blocks (LCBs)
- 2.86Mbp of *backbone* sequence – only 58% of average genome size
- *Backbone* is any region shared among all genomes



File Input/Output to mauveAligner

- Input either GenBank or Multi-FastA sequences

  - Incomplete genomes are OK

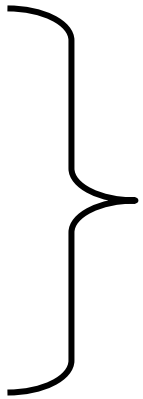
- Output alignment in eXtended Multi-FastA (XMFA)
- Output phylogenetic tree in Newick standard format
- Output a list of island and backbone locations

File Input/Output to mauveAligner

An example of the XMFA format for two sequences. XMFA is also used by Shuffle-LAGAN.

```
>1:325-5000 + S_typhi.gbk
AC-TG-NAC--TG
AC-TG-NACTGTG
...
>2:7675-3000 – E_coli.gbk
AC-TG-NAC--TG
AC-TG-NACTGTG...
=

>1:5500-6100 + S_typhi.gbk
AC-TG-NAC--TG
AC-TG-NACTGTG
...
>2:600-1000 + E_coli.gbk
AC-TG-NAC--TG
AC-TG-NACTGTG
...
=
```



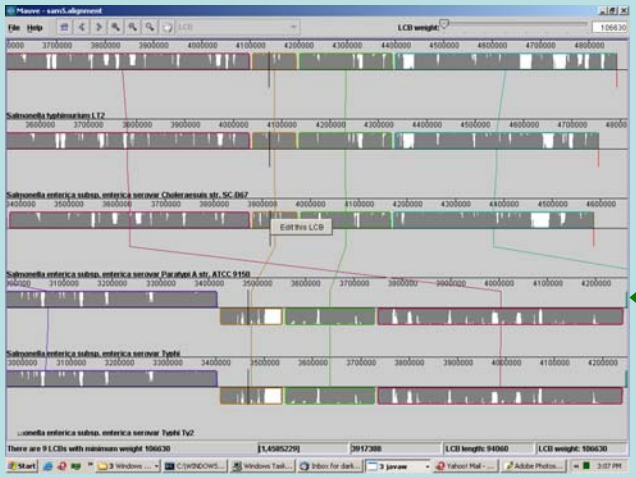
Each Locally Collinear Block gets an entry like this, separated by the = character.

A collection of LCBs constitute the genome rearrangement structure

More details are available in the Mauve Documentation online at **http://gel.ahabs.wisc.edu/docserver/mauve**

The two component architecture of Mauve

Java 1.4 Interactive Visualization

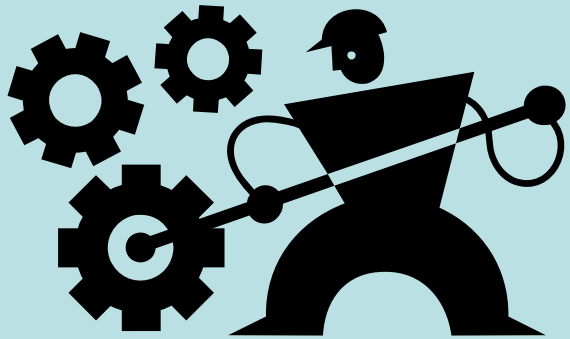


GenBank or FastA sequences

alignments

C++ command-line aligner

Windows, Linux, Mac OS X

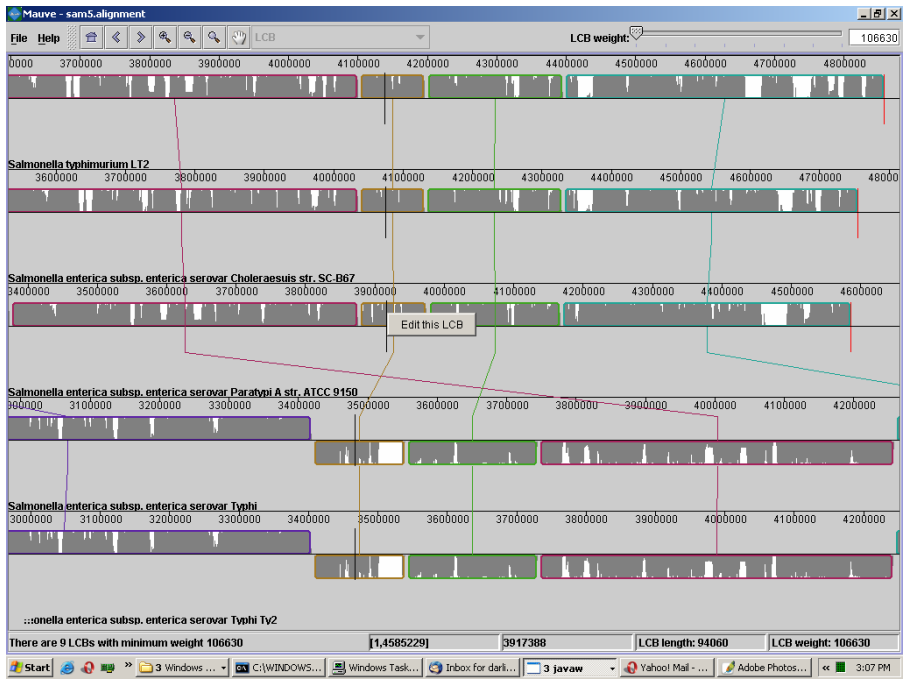


100% Free/Open Source Software

We use each language for what it does best–C++ for efficient algorithm implementation, Java for a cross platform GUI

The Mauve visualization system

- We want to visualize rearrangement structure in genomes
- Previous comparative genome browsers either did not display rearrangements (Vista, GBrowse) or handled them poorly, not showing breakpoints or similarity profiles (ACT, GenomeViz, GATA)
- A screenshot of Mauve:



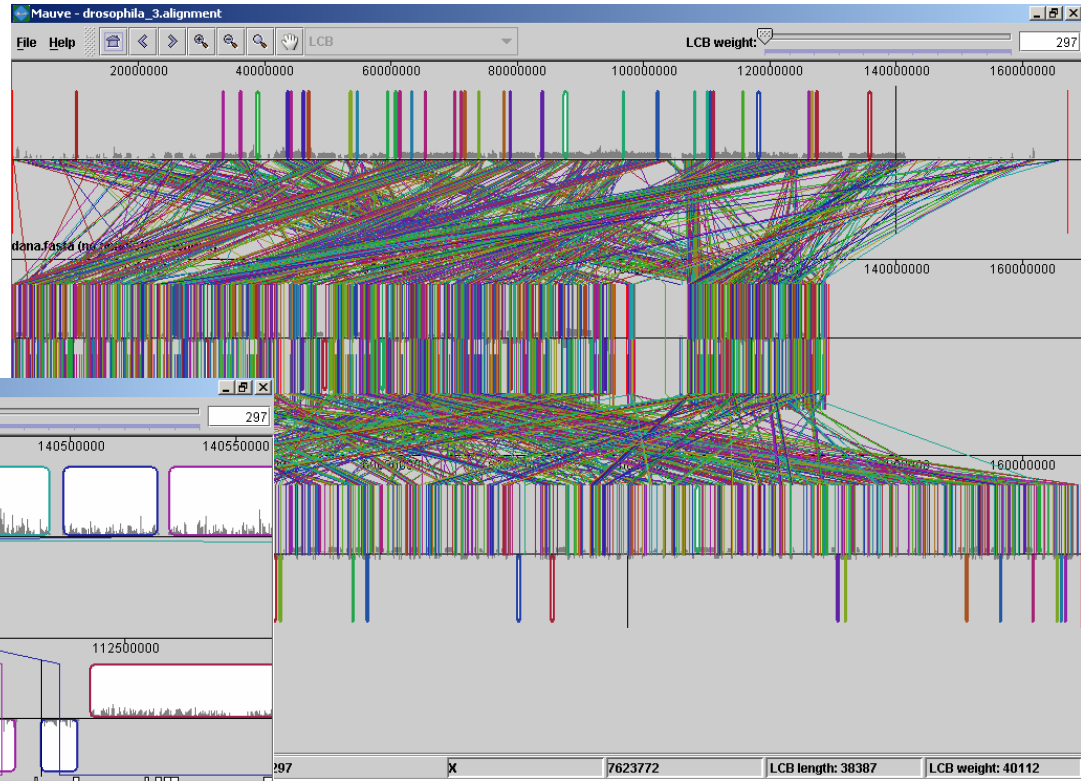
Visualization Design goals

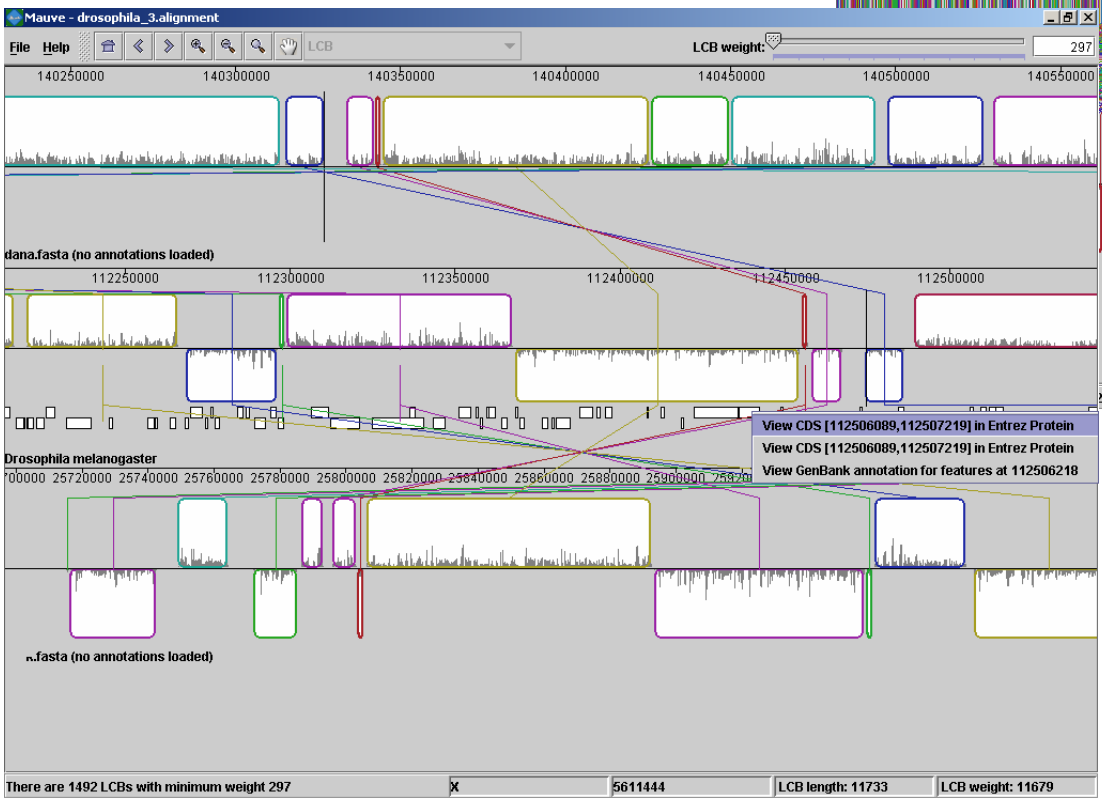
- Display new types of data about genome structure
  - Rearrangements, differential content
- Interact with large genomic data sets gracefully
  - Other tools are unacceptably sluggish
- Modular and extensible code
  - Event-driven, data model separation, JavaDoc
  - Mauve can be tailored to the task at hand
- Leverage existing code wherever possible
  - Use BioJava to read data and display annotation
  - Edit alignments with Cinema-MX and (soon!) BaseByBase
- Support both stand-alone use and integration with web-accessible databases

Mauve viewing 3 *Drosophila* genomes

A global map of rearrangement

A detailed view of the X chromosome

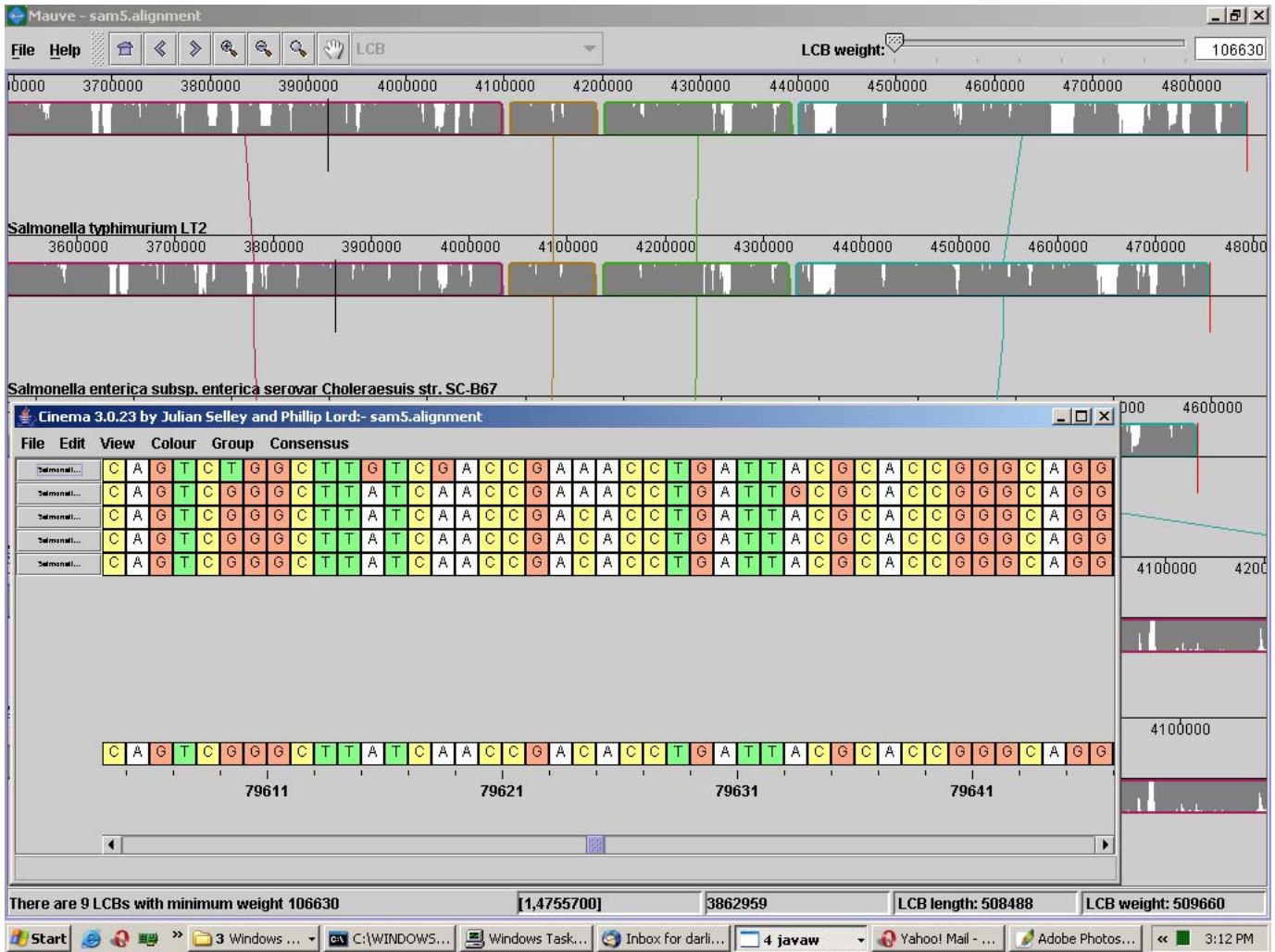




The popup menu launches a web browser to a cross-referenced database. Any web based DB with a GenBank db_xref will work

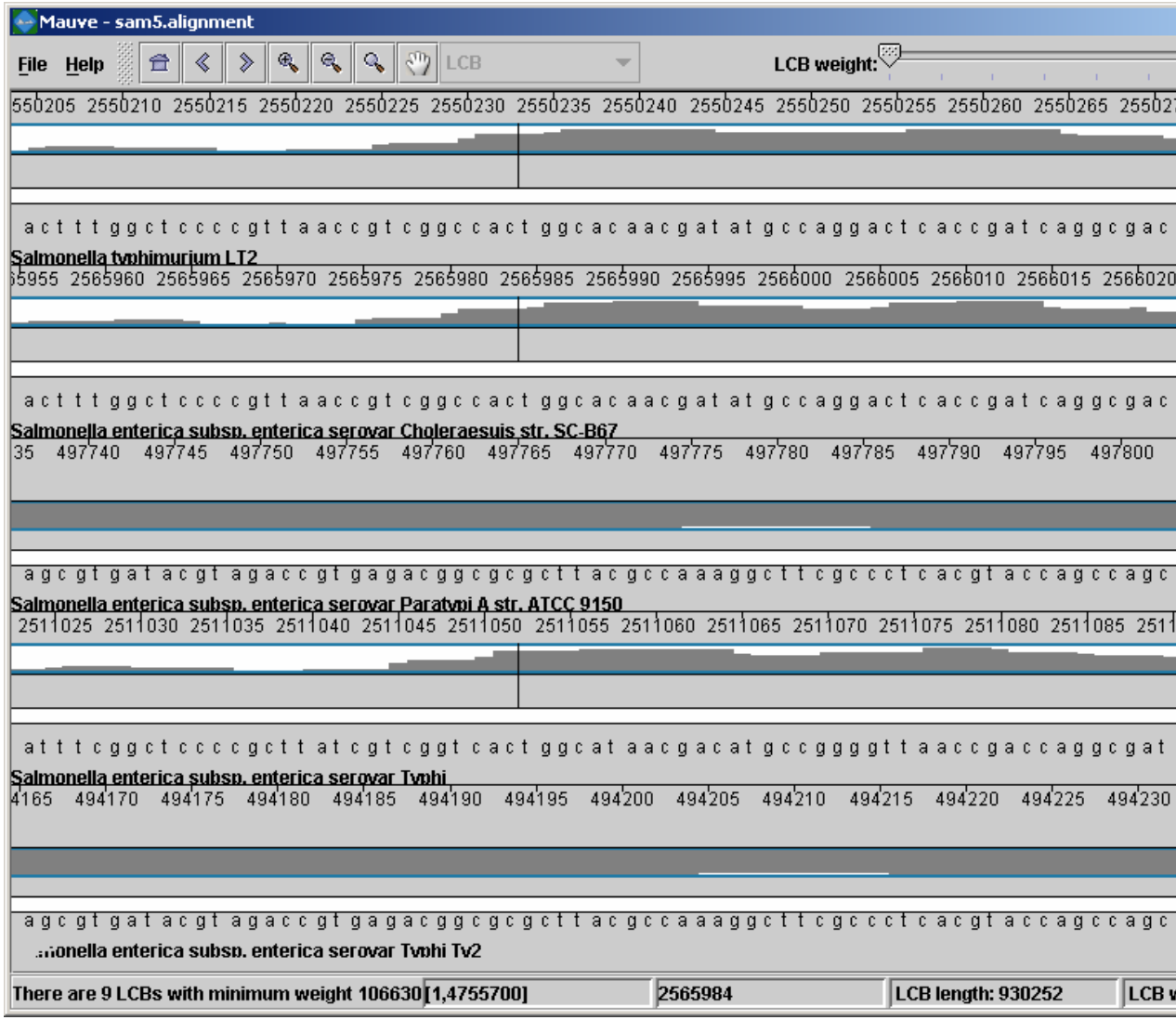
Preliminary alignment editing support

By right-clicking an LCB, an option to edit the alignment can be selected…



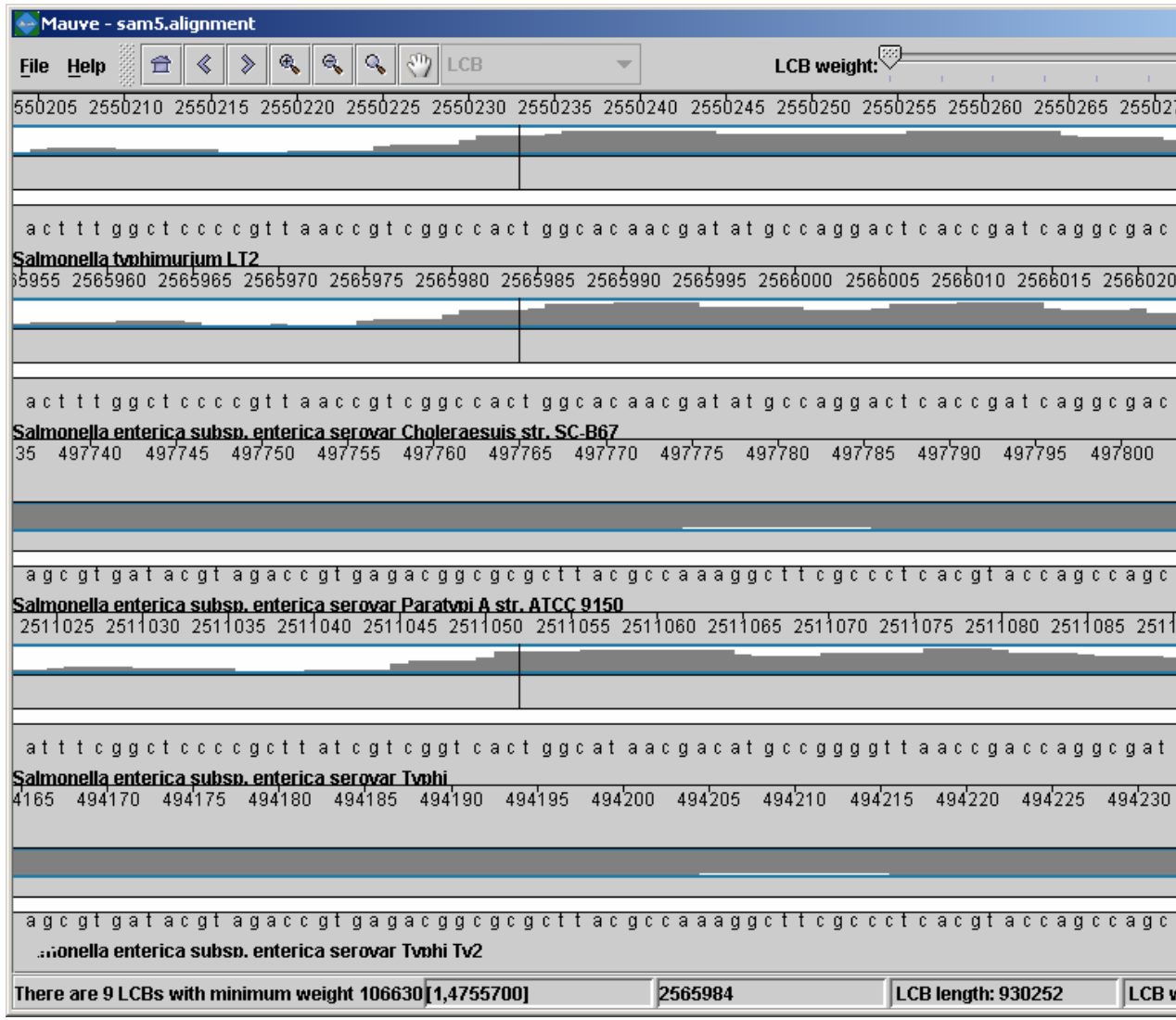
The anatomy of the display…

The sequence display is contained by a **RearrangementPanel**



The anatomy of the display…

Each genome is contained by a **SeqPanel**



The anatomy of the display…

Each SeqPanel contains...

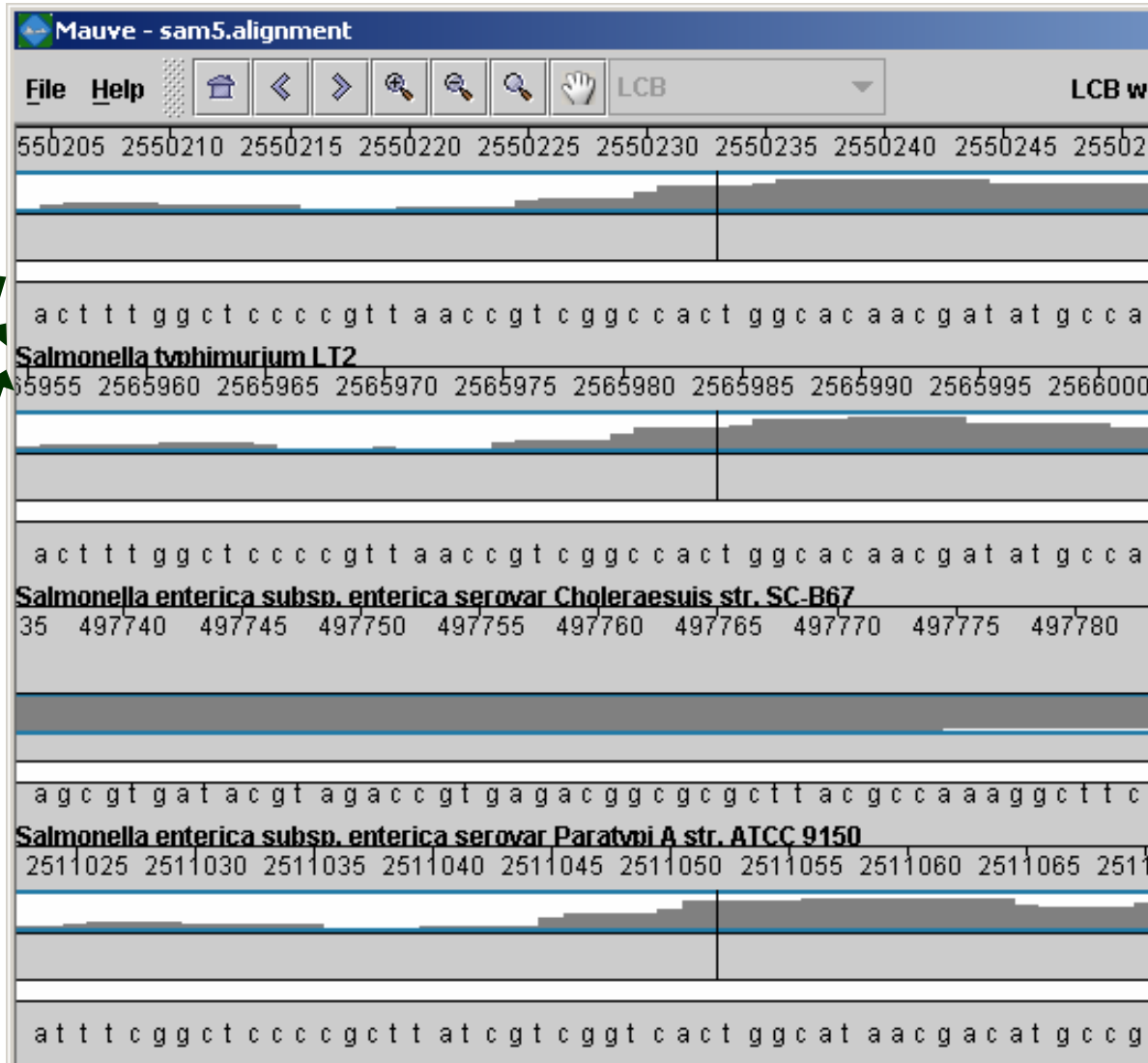
A BioJava Ruler Panel

An RRSequencePanel (for the similarity profile)

A BioJava FeaturePanel (for annotated features)

A BioJava SymbolList (the nucleotide sequence)

The organism name



Deploying Mauve

Scenario (1) Stand-alone

- Java visualizer packaged with C++ aligner
- Alignments run locally on user's computer

Scenario (2) web service

- Sequences are submitted for alignment via a web site
- The alignment job runs on a compute cluster
- Mauve uses Java Web Start to launch a result display

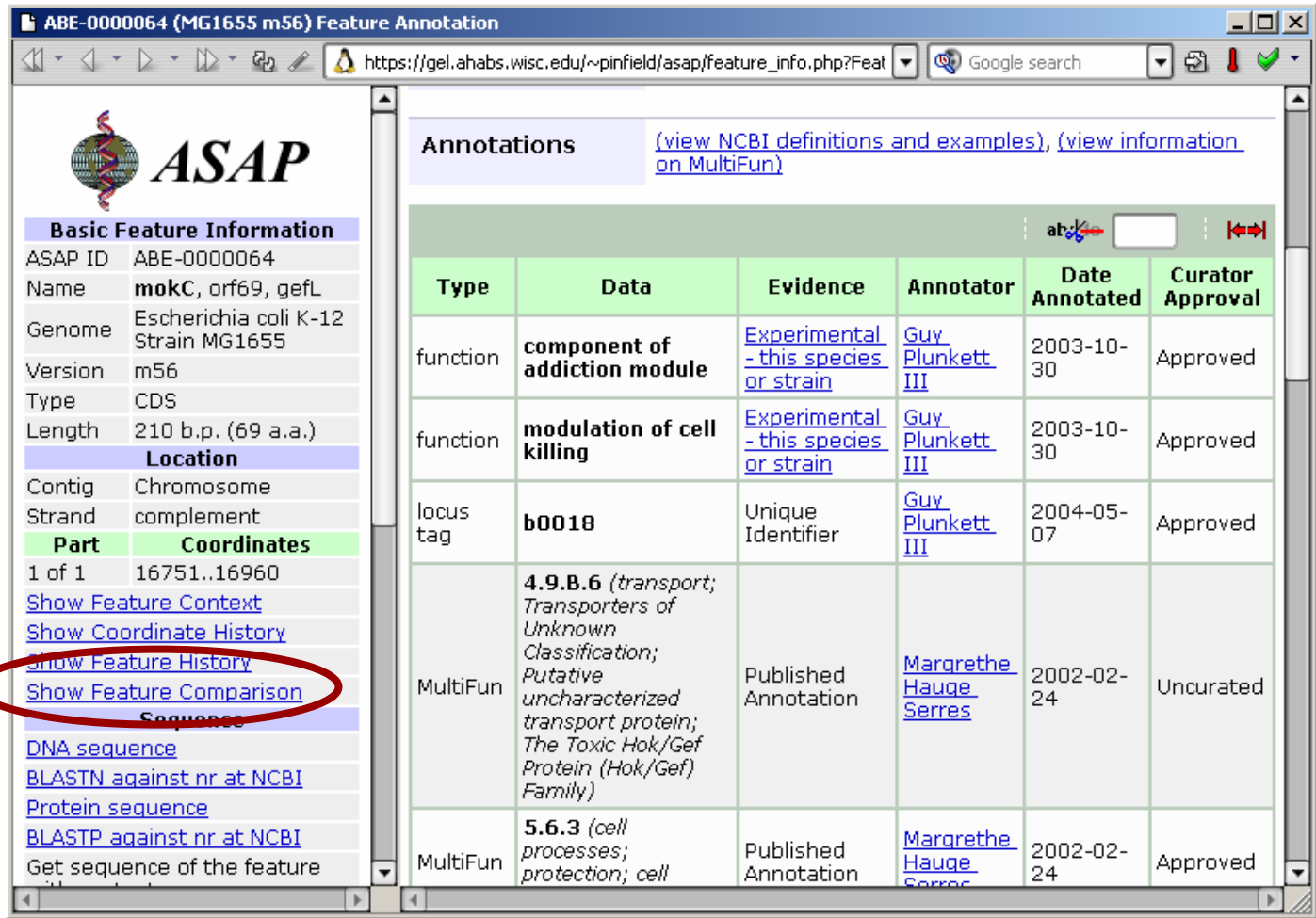
Scenario (3) database integrated

- A set of alignments (possibly curated) exist in the database
- When viewing a feature, the user can launch Mauve with alignments containing that feature

Database Integrated Mauve…

When “Show Feature Comparison” is clicked, a Mauve window is launched with a multiple genome alignment.

If Mauve is already running, the display shifts to show the feature.



Mauve and ERIC

Precomputed Alignments for Complete Genomes:

- *E. coli and Shigella*
- *Salmonella*
- *Yersinia*
- *E. coli, Shigella, Salmonella*
- *All organisms?*

Precomputed Alignments for Draft Genomes:

- *within Genera*

Developer Information

Source code is maintained in a Subversion repository

Nightly snapshots are available at
http://gel.ahabs.wisc.edu/mauve/source/snapshots/

Build environment

Eclipse and Apache Ant for Java

Visual Studio 2005 and gcc 4.0.0 with autotools

API documentation

JavaDoc and Doxygen format

User's guide served by a Zope BackTalk server

Bug Tracking and Feature Requests handled by a Mantis server.

All of this information is available at
http://gel.ahabs.wisc.edu/mauve/developer.php